DOCUMENT RESUME

ED 472 522 TM 034 588

AUTHOR	Dings, Jonathan; Childs, Ruth; Kingston, Neal
TITLE	The Effects of Matrix Sampling on Student Score Comparability in Constructed-Response and Multiple-Choice Assessments. Technical Guidelines for Performance Assessment.
INSTITUTION	Council of Chief State School Officers, Washington, DC.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	2002-00-00
NOTE	34p.; Produced by the State Collaborative on Assessment and Student Standards.
CONTRACT	R279A50006
AVAILABLE FROM	Council of Chief State School Officers, One Massachusetts Avenue, Suite 700, Washington, DC 20001-1431 ($10). Tel: 202-408-5505; Fax: 202-408-8072. For full text: http://publications.ccsso.org
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	*Junior High School Students; Junior High Schools; *Performance Based Assessment; Reliability; *Sampling; *Scores; *Test Construction; *Test Content
IDENTIFIERS	*Multiple Matrix Sampling

ABSTRACT

This study examined matrix sampling of test content, the practice of giving various students in the same school differing test questions. This often-used approach to large-scale assessment allows for relatively broad coverage of the curriculum, but with fewer comparable individual student scores than a conventional test. One can be sure that scores will differ, but no known research tells how large the differences will be. The study used a quantitative research technique to study how much the comparability of individual student scores is sacrificed when tests are matrix sampled. Tests of varying lengths, comprising multiple-choice items, constructed-response items, and a combination of both were examined. Student scores on "pseudo-forms" were studied using forms of a grade 8 statewide mathematics test, with more than 4,000 students taking each of 12 test forms. Results show that matrix sampling reduced the comparability of student scores on average by about one-fifth to one-third of a standard deviation unit when a small percentage of the items were matrix sampled, and by about six-tenths to more than four-fifths of a standard deviation unit when the entire test was matrix sampled. Results varied according to the reliability of the test scores studied, and do not necessarily generalize to other test lengths or to data from other testing programs. The study illustrates a technique test developers could use to see how greatly the comparability of individual student scores is compromised by a matrix sampling test design. An appendix shows the percent of items covering content across the 12 original test forms. (Contains 8 tables and 18 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 472 522

TM034588

SCASS

State Collaborative on Assessment and Student Standards

SCASS TGPA

Technical Guidelines for Performance Assessment

The Effects of Matrix Sampling on Student Score Comparability in Constructed-Response and Multiple-Choice Assessments

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

B. Buterbaugh

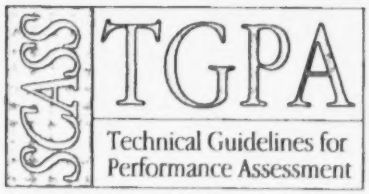
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

BEST COPY AVAILABLE

Jonathan Dings
Ruth Childs
Neal Kingston

A Publication of the Council of Chief State School Officers



The Effects of Matrix Sampling on Student Score Comparability in Constructed-Response and Multiple-Choice Assessments

JONATHAN DINGS, BOULDER VALLEY PUBLIC SCHOOLS

RUTH CHILDS, ONTARIO INSTITUTE FOR STUDIES IN EDUCATION
OF THE UNIVERSITY OF TORONTO

NEAL KINGSTON, MEASURED PROGRESS

January 2002

Council of Chief State School Officers
One Massachusetts Avenue, N.W.
Washington, DC 20001

This report was prepared for submission under contract with the Council of Chief State School Officers and funded by the United States Department of Education, Office of Educational Research and Improvement, Grant No. R279A50006. The views and opinions expressed in this report are not necessarily those of the United States Department of Education, the Council of Chief State School Officers, the State Collaborative on Assessment and Student Standards, or the states participating in the study.

COPYRIGHT © 2002 BY THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS, WASHINGTON, DC

The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues. Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center was established through a resolution by the membership of CCSSO in 1984. This report is sponsored by the Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA) consortium. The SCASS TGPA works with researchers to design and implement practical and timely research on large-scale performance assessment. This research provides information useful in designing state assessment and accountability programs so that they yield results that can be used to improve student learning.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

G. Thomas Houlihan
Executive Director

Wayne N. Martin
Director
State Education Assessment Center

John F. Olson
Director of Assessments
State Education Assessment Center

Phoebe C. Winter
Project Director, Technical Guidelines for Performance Assessment
State Collaborative on Assessment and Student Standards

ACKNOWLEDGMENTS

This project was conducted with the assistance of members of CCSSO's State Collaborative on Assessment and Student Standards (SCASS) Technical Guidelines for Performance Assessment (TGPA). We thank members of the study group overseeing this project: James Masters of the Pennsylvania Department of Education, Phoebe Winter of the Council of Chief State School Officers, and Liru Zhang of the Delaware Department of Public Instruction. We also thank two reviewers whose comments were instrumental in shaping this study: Anne Fitzpatrick of CTB-McGraw Hill, and Edward Haertel of Stanford University. Any errors or omissions remain our own.

CONTENTS

EXECUTIVE SUMMARY 1

PROVOCATION 3

BACKGROUND 4

RESEARCH QUESTION 5

METHOD 7

- Overview 7
- Data 7
- Creating Pseudo-Forms 7
- Rescaling Pseudo-Form Scores 8
- Technical Details of the IRT Scaling 8
- Multiple-Choice and Constructed-Response Score Combinations Studied 9
- Comparisons Made 9
- Degrees of Matrixing Studied 10

RESULTS 11

IMPLICATIONS 13

LIMITATIONS AND EXTENSIONS 15

REFERENCES 17

TECHNICAL NOTE 19

FIGURES AND TABLES 20

- Figure 1: Average Difference in Scores Between Pseudo Forms in Standard Error of Measurement Units (RMSD/SEM), Plotted Against Percent of Items in Common Across Pseudo-Forms 20
- Figure 2: Average Difference in Scores Between Pseudo-Forms in Standard Deviation Units (RMSD/SD), Plotted Against Percent of Items in Common Across Pseudo-Forms 22
- Table 1: Ways of Combining Six 5-Item Sets into Pseudo-Forms of 3, 4, or 5 Sets Each 24
- Table 2: Median Summary Statistics for Original and Pseudo-Form Raw Scores, Computed Over 12 Test Forms of Multiple-Choice and Constructed-Response Items 25
- Table 3: Median Standard Deviation (SD) and Standard Error (SE) for Item Response Theory Based Scale Scores on Original and Pseudo-Forms 26
- Table 4: Pseudo-Form Lengths Employed to Study Various Combinations of Multiple-Choice and Constructed-Response Items, Together and Separately 27
- Table 5: Median Root Mean Squared Difference and Pearson Correlation Between Pseudo-Form Scores, with Reliability Estimates and Number of Replications 28
- Table 6: Average Difference in Scores Between Pseudo-Forms in Standard Error of Measurement Units (RMSD/SEM), by Percent of Items in Common Across Pseudo-Forms 29
- Table 7: Average Difference in Scores Between Pseudo-Forms in Standard Deviation Units (RMSD/SD), by Percent of Items in Common Across Pseudo-Forms 27
- Table 8: Example of Generalizability Study Results, Form 1, Design With 80 Percent of Items in Common Across Pseudo-Forms (500 Students, 2 Scores per Student) 29

APPENDIX A 29

Percent of Items Covering Content Across the Twelve Original Test Forms, By Constructed-Response versus Multiple-Choice Item Format 29

EXECUTIVE SUMMARY

In this report, we addressed matrix sampling of test content—the practice of giving various students in the same school differing test questions. This often-used approach to large-scale assessment allows for relatively broad coverage of the curriculum, but with less comparable individual student scores than a conventional test. Because students do not all answer exactly the same questions, their scores will differ based on individuals' knowledge of the particular content tested by the questions they receive. Although we can be sure that scores will differ for this reason, no known research tells us how large the differences will be.

We employed a quantitative research technique to study how much the comparability of individual student scores is sacrificed when tests are matrix sampled. We examined tests of varying lengths, comprised of multiple-choice items, constructed-response items, and combinations of the two. We found that matrix-sampling reduced the comparability of student scores on average by about one-fifth to one-third of a standard deviation unit when a small percentage of the items were matrix sampled, and by about six-tenths to more than four-fifths of a standard deviation unit when the entire test was matrix-sampled. Results varied according to the reliability of the test scores studied and do not necessarily generalize to other test lengths or to data from other testing programs.

With this study, we have illustrated a technique test developers could use to see how greatly the comparability of individual student scores is compromised by a matrix sampling test design. By contrast, policymakers and other potential users of matrix sampled tests should consider our findings only as a very rough indication of the loss in comparability that might occur with matrix sampling of tests similar to the one we examined, not as a definitive prediction of results that will be obtained for a particular testing program.

PROVOCATION

Imagine a district, province, state, or nation that undertakes to develop a testing program for school accountability. Policymakers interpret the laws or regulations and, most likely, craft a request for proposals, and prospective contractors bid to develop the test. Between the conception of the program and the time a test booklet hits students' desks (or items appear on their computer screens), policymakers, administrators, technicians and/or the contractor design a system of tests to measure implementation of the target curriculum. The structure of each test will reflect prioritization of two competing needs: the need to ensure that each student answers questions that are equivalent in content to those posed of every other student; and the need to cover the curriculum broadly, with a sufficient number of questions to minimize ill effects of teaching to the test, if not teaching the test.

Whenever student accountability is the primary purpose, as with a test used to determine eligibility for graduation, the need to administer each student an equivalent test will take precedence over the need to cover the curriculum broadly. Most likely, only a single test form will be administered at each test administration, and test content across administrations will be matched at a very detailed level across statistically parallel test forms. Alternatively, only a common set of items embedded in each of multiple forms will count toward student scores. Either way, curriculum coverage will need to be limited to the single form and its close cousins, if the graduation testing program is to have much of a chance of surviving legal challenges.

When, instead, school accountability is the primary purpose of testing, informed policymakers must consider how to ensure broad enough content coverage that teaching to the test can only be accomplished through careful attention to teaching the entire curriculum. Doing so will require either that each student tests for an impracticable length of time, or that students in the same classroom take different test forms constructed, as a set, to cover the curriculum by dividing it into parts that will be covered on only a few or even just one of several forms. By contrast to the single form approach mentioned in the context of the graduation test, this approach is known as matrix sampling of test content.

If asked by policymakers to compare competing needs for broad curriculum coverage and giving all students tests that cover equivalent content, testing contractors could give many practical reasons to prefer a single-form design. The single-form design is easy to explain to policymakers and minimizes costs associated with item development, test printing, scoring, and, for the most part, technical work (psychometrics) necessary to generate results that can be compared from year to year. However, beyond observing that individual student results are not genuinely comparable to each other unless forms are built to the same specifications, there is scant published quantitative evidence to illuminate just how much matrix sampling compromises the comparability of individual student scores. In this report, we review what is known about the effects of matrix sampling and demonstrate a quantitative research technique that could be used to inform policymakers and potential users of matrix sampled tests about the loss in comparability brought about by matrix sampling.

BACKGROUND

Matrix sampling is one of the most widely used approaches to providing broad content coverage in large-scale assessment programs. This approach involves developing a complete set of items judged to cover the content area, then administering each student a subset of the items. Matrix sampling, by limiting the number of items administered to each student, limits the amount of testing time required, while still providing coverage of a broad range of content across students.

Although use of matrix sampling in test development has been noted as a potentially economical means of estimating item parameters (Garg, Boss, & Carlson, 1986), assessment programs typically use matrix sampling designs to produce school, district, state, or national results. For example, the National Assessment of Educational Progress (NAEP) uses a matrix sampling design in which the complete set of items is divided into non-overlapping blocks, which are then combined into booklets in such a way that every block is paired with another block. This spiraling design minimizes the testing burden for individual students but still permits estimation of proficiency distributions for states and the nation using plausible values methodology (Mislevy, Beaton, Kaplan, & Sheehan, 1992). NAEP is not used to produce results for individual students, however.

Indeed, most of the literature on matrix sampling concerns the production of aggregate results. Lord (1962) describes the estimation of item statistics for groups of students. Bock and Mislevy (1981) describe analytic techniques that produce school-level results, but not results for individual students. By contrast, Bock and Mislevy (1987) also propose a "duplex design" that, through careful content sampling in developing the subsets of items, might provide both school-level results for broad content coverage and individual-level scores. However, this design requires such a large item pool for test development as to be impractical (Kerins & Brennan, 1987). Thus, we must consider alternative ways of implementing matrix sampled test designs.

Matrix sampling, as traditionally applied, addresses the concern about content coverage but limits the comparability of individual scores. Two students who were equally knowledgeable about some topics in a content area would probably receive quite different scores if one received a test on those topics and the other received a test on topics neither student had yet learned about. The other traditional approach to achievement testing —administering a single set of common items to all students — will produce the highest level of comparability because all the students have the same opportunity to answer exactly the same questions. A compromise solution that is currently being used or has recently been used in a number of state testing programs (e.g., Kentucky, Maine, Massachusetts, and New Hampshire) is a "partial" matrix sampling design in which some items are administered in common to all students and other items are matrix sampled. This approach uses a single test design to achieve the goals of providing adequate content sampling *and* providing adequately reliable and reasonably comparable scores for individual students, within the constraints of administering a given number of items to each student. This design offers testing programs some insurance against a substantial concern—narrowing the curriculum—that is heightened by use of a single test form, as well as a larger operational item pool to meet the recognized need to release items annually for public use (Linn, 2000).

RESEARCH QUESTION

Although several researchers have examined the issue of comparability of performance assessment results at the individual level or in aggregated form (e.g., Brennan & Johnson, 1995; Cronbach, Linn, Brennan, & Haertel, 1995; Fitzpatrick & Yen, 1999; Haertel & Linn, 1996; Yen, 1997), no known previous study has attempted to quantify the degree to which matrix sampling reduces the comparability of student scores. In this study, we evaluate the levels of statistical comparability in individual student scores arising from partially matrix sampled test designs representing different ratios of common items to matrix items, including a completely matrix-sampled design. Because many testing programs incorporate both multiple-choice and constructed-response items, we compare the effects of matrix sampling on score comparability for differing item types, employing data from a state testing program.[1]

[1] The state cannot be named in this report under terms of the research agreement enabling this study.

METHOD

OVERVIEW

To study the error introduced by matrix sampling, we compared student scores on "pseudo-forms," shortened test forms created from an existing test. At its heart, our strategy was to compute two scores for each student, each different from the other and from the total score on the existing test; the two scores were based on some, but not all, of the items taken by that student. Because the two scores were not both calculated from all the same test items, we analyzed statistical differences between the two scores to evaluate score comparability. We repeated our procedure on many different combinations of items for each student and for varying combinations of multiple-choice and constructed-response items, hoping to find a pattern of results to describe the effect of matrix sampling on student score comparability.

DATA

Forms of a grade 8 statewide mathematics test were used as the source of data for pseudo-forms in this study. Forms included multiple-choice and extended constructed-response items, targeted to cover seven content categories: Number; Mathematical Procedures; Space and Dimensionality; Measurement; Change; Mathematical Structure; and Data. For relative proportions of the content covered, see Appendix A. Note that the multiple-choice and constructed-response items covered somewhat different content. We label the constructed-response items as "extended" to indicate they were frequently multiple-part items, requiring explanation and/or extension of obtained results to the general case.

Twelve forms of the test were administered to nearly all eighth-graders in the state, with over 4,000 students taking each form. The forms were spiraled within school for administration, meaning that booklets were distributed in sequence (that is, ordered in a repeating pattern from one to twelve) to establish a sequentially stratified, randomly equivalent sample of students taking each form. Each form consisted of 6 constructed-response and 24 multiple-choice items, not counting pre-test items, which were ignored in our study. The twelve forms were partially matrix sampled, with some items in common across the forms. This did not affect our analyses, because the results were computed separately for each form.

CREATING PSEUDO-FORMS

We considered each of the 12 original forms separately, as though each form were made up of six sets of five items, each set containing one constructed-response item and four multiple-choice items, as shown in Table 1. The four multiple-choice items in each item set were stratified by position in the test booklet, with every sixth one contributing to an item set. The first item set contained the first, seventh, thirteenth and nineteenth multiple-choice items on the original form, as well as the first constructed-response item; the second item set contained the second, eighth, fourteenth and twentieth multiple-choice items on the original form, as well as the second constructed-response item; and so forth. Stratifying by position in this way helped to minimize the contribution of

possible context and occasion effects to differences across item sets, as the original test was not necessarily administered in a single sitting (it was designed to be administered in up to three sessions totaling two and one-half hours).

We combined item sets to create pseudo-forms—collections of 3, 4, or 5 sets of five items (thus, containing 15, 20, and 25 items, respectively), which reflected each student's performance on fewer than the total of the six 5-item sets (30 items) that make up the whole test form. As shown in Table 1, we included all possible combinations of 3, 4, or 5 sets of items in our study. Raw-score statistics in Table 2 summarize psychometric properties of the original form, consisting of all 6 five-item sets, and each pseudo-form. As expected given the forms' construction, noticeable differences are apparent among mean scores of several pseudo-forms of the same length, with somewhat less apparent differences in score standard deviations and reliabilities.

Rescaling Pseudo-Form Scores

Having identified differences in scores as shown in Table 2, we employed a commonly used scaling technique capable of virtually standardizing score means and standard deviations across pseudo-forms. Had we not standardized in this way, comparisons of scores on any pair of pseudo-forms would always have reflected the relative difficulty of their constituent item sets. Students would have received, on average, higher scores on easier pseudo-forms, yielding greater mean differences between an easy and a difficult pseudo-form than between two of comparable difficulty.

To focus score comparisons as directly as possible on differences in content covered by pseudo-forms, we scaled student scores using techniques based on Item Response Theory (IRT). Specifically, we employed a two-parameter partial credit model for constructed-response questions, as scored on a five-point scale (0 to 4), and a three-parameter logistic model for multiple-choice questions. This yielded mean scores of zero (when rounded to the hundredths place) for every pseudo-form. In Table 3, we present score standard deviations and standard errors of estimate, a measure of the degree of random variation expected in IRT-based scores that increases as score reliability decreases. Both statistics vary only slightly across pseudo-forms of the same length, with generally smaller differences across multiple-choice scores than constructed-response. We present correlations between multiple-choice and constructed-response scores in Tables 2 and 3 as a way of showing both the high degree of statistical relationship between the two item types, and that the degree of relationship between the types for IRT scale scores is equal to or marginally greater than the relationship between raw scores. Both tables summarize findings across the dozen forms used in the testing program. As with the remainder of this study, all analyses were completed on each test form of the eighth grade mathematics test separately and combined in analyzing and interpreting our results.

Technical Details of the IRT Scaling

Item parameters and student scores were estimated using a large data set containing a record of the original responses for each student, as well as additional records for student responses on each pseudo-form. For example, for pseudo-forms three sets of items long, each student had 21 response records in the data set, one for the total base form and one for each of the 20 possible pseudo-forms that could be constructed. Item sets not included on a particular pseudo-form were treated as "not presented" in scaling with PARSCALE (version 3.2, Muraki & Bock, 1996). In all, student scores were calculated

for 41 pseudo-forms[2]. Expected *a posteriori* estimation was used to compute the student score estimates.

Because linking scores across the twelve forms used in the state testing program were not necessary to this study, item parameters and student scores were estimated separately for each test form. The large sample size, approximately 3,200 students per form, supported this approach. These students were selected from those who demonstrated an effort to take the test seriously (for example, did not leave all constructed-responses blank) and made up about eighty percent of that state's eighth graders.

Multiple-Choice and Constructed-Response Score Combinations Studied

PARSCALE results included separate student scores calculated from multiple-choice and constructed-response questions, as well as linearly combined scores reflecting relative weights we selected to study differences between the performance of constructed-response and multiple-choice questions. We chose an equal weighting of scores from the two item types, as well as combinations reflecting two-to-one ratios of multiple-choice to constructed-response scores and of constructed-response to multiple-choice. These weighting schemes are options likely to be considered for testing programs in weighting constructed-response and multiple-choice questions that yield the same total number of points, as do the sets in our pseudo-forms. Similarly, IRT scaling is a likely option for state testing programs to scale together multiple test forms with only some items in common.

Comparisons Made

After producing student scores on each pseudo-form, we compared correlations, generalizability analysis results (Brennan, 1983), and Root Mean Squared Differences (RMSDs) across pairs of pseudo-forms. RMSD is a measure of the average absolute difference between scores of each student on two different pseudo-forms, computed by squaring the difference between each student's score on, for example, pseudo-form 3A and pseudo-form 3B, averaging these squared differences across all students, and taking the square root of the average. RMSD provides some indication of how far the results deviate from the theoretical ideal (in this case, a correlation of unity across scores displaying the same mean, standard deviation, etc.). Correlations are included for descriptive purposes.

For computational efficiency, we randomly selected scores of 500 students per pseudo-form pair to use in analysis of variance-based generalizability estimates. Trimming the sample to this size provided a reasonable estimate for each form, while still generating results based on 6,000 students across the twelve forms. A non-trivial saving of computer run time resulted, considering the number of times each analysis was performed (described below).

For any comparison of pseudo-forms listed in Table 1, each student had provided responses for all the items on both pseudo-forms, enabling a one-facet generalizability analysis of students crossed with pseudo-forms, studied using IRT-based scores computed for every student on each pseudo-form. Generalizability analysis yields estimates of universe score variance (i.e., variance of scores due to students) and two

[2] Users of IRT-scaling software programs will note alternative ways of accomplishing the same purpose: calibrating items based on a given data set, and computing examinee scores using another.

kinds of error variance: variance due to pseudo-form and a residual error encompassing variance due to the student by pseudo-form interaction.

DEGREES OF MATRIXING STUDIED

Comparing a student's score on one pseudo-form with his/her score on another using the statistics noted above indicated the difference in performance attributable to the item sets that were not shared across the pseudo-forms. Continuing our example, pseudo-forms 3A and 3B are three sets of 5 items long. As show in Table 1, they have two item sets in common. Thus, they share 67% of their length. Scores from 3A and 3H are based on one-third the same items, whereas 3A and 3T have none in common. By considering several ratios of common items to matrix items, including those derived from pseudo-forms that have four and five sets of items, we hoped to observe a pattern in the results.

Table 4 details the comparisons made among pseudo-forms used to model matrix sampling situations where 80, 75, 67, 50, 33, and 0 percent of items were common across forms. The first entry shows a pseudo-form length of five item sets (25 items, of which 5 are constructed-response and 20 are multiple-choice). Fifteen separate pairs of pseudo-forms of this length were compared, with 80% of items in common across each pair. Note that pseudo-form length is not crossed with percent of items in common in our design. For example, 80 percent of items are in common only for pseudo-forms that are five sets of items in length. Pseudo-forms that are 4 sets of items in length provide our only comparisons with 50 and 75 percent of items in common.

We also note that not all possible pairs of pseudo-forms were studied in every case, because this would have greatly increased the number of replications necessary to complete the study. Although all 15 of the possible pairings of the 5 sets of items or 80% of the items in common condition described above were studied, only 17 of the 79 possible pairings in the 3 sets of items or 67% of the items in common condition were included. Lest this latter number seem small, we hasten to point out that each analysis was conducted on twelve test forms, resulting in 204 replications in total for the 67% of items in common condition. In all, analyses were completed for 136 pairings of pseudo-forms per original form, a total of 1632 (12 x 136) pairs, with 5 scores per pseudo-form: constructed-response only, multiple-choice only, and the three weighted composite combinations of constructed-response and multiple-choice scores.

RESULTS

Not surprisingly, comparability of scores declined as fewer items were held in common across pairs of scores that were compared for each student. Stated another way, the error introduced by matrix sampling increased with the degree of matrixing. RMSD values in Table 5 indicate a greater average absolute difference between pseudo-forms scores based on no item sets in common than in cases where one-third of item sets were in common, which were in turn greater than cases where half were in common, etc. This pattern held for scores based on CR or MC items and all combinations thereof. RMSDs were larger for CR than MC scores, probably reflecting higher MC score reliability. Pseudo-form score correlations followed the pattern of the RMSDs: Student scores were correlated more highly across pseudo-forms when more of the item sets were in common.

By themselves, correlations and RMSD statistics do not provide a completely adequate context for judging how substantial a score difference is. To control for differences in pseudo-form length and to make sense of results relative to score reliability, we divided our indicator of score difference by the average standard error of measurement (SEM) for the pseudo-forms being compared. SEM was calculated by squaring the asymptotic standard error of estimate of ability for each examinee, and taking the square root of their average for each pseudo-form. Results (Table 6, plotted in Figure 1) indicate a not quite linear relationship between percent of items in common across pseudo-forms and RMSD/SEM, which is the amount of error introduced by matrixing, relative to test reliability. For example, having only one-third of items in common across pseudo-forms in our test data introduced nearly one standard error of measurement difference into comparisons of student scores across matrixed forms, whereas having none in common added about 1.15 SEM to such score comparisons. Interestingly, there were negligible differences in this regard between MC and CR results, marginally favoring CR.

Despite the simplicity of a consistent finding for RMSD/SEM figures across different linear weighted combinations of MC and CR items, it is somewhat difficult to interpret average absolute score differences relative to SEM alone. Results in Table 6 and Figure 1 suggest, for example, that test developers must pay a fairly dear price in terms of student score comparability to matrix even one-fifth of the items an examinee takes: Just over one half a SEM unit. By contrast, matrixing the entire test requires merely a little more than twice the loss in score comparability: About 1.15 SEM. To place the loss in comparability in the context of the scores used, we present in Table 7 and Figure 2 RMSDs standardized against observed score standard deviations. In this analysis, the more reliable the score, the smaller the loss in standard deviation units. Matrixing one-fifth of the test results in losses in score comparability from one-fifth to one-third of a standard deviation unit. Matrixing the entire test yields losses approximately three times greater.

For these data, correlations were identical to results of generalizability analyses, to the number of decimal places given in Table 5. This finding reflects IRT scaling having removed mean differences (main effects) from pseudo-form comparisons, a property that is typically advantageous in scaling matrix sampled tests together and is documented for these data in Table 3. Generalizability data are not presented for each form to avoid redundancy with the correlations in Table 5. As an example of our calculations,

however, we present results from a comparison of pseudo-form scores on a single test form in Table 8. Scores in this example reflect an equally weighted combination of multiple-choice and constructed-response results. Degrees of freedom, sums of squares, and mean squares are the more familiar result of analysis of variance procedures, whereas the estimated variance components and percent of total variance accounted for are from generalizability analyses. Generalizability analysis results were obtained through standard computational formulas using mean squares and numbers of examinees and pseudo-forms to compute variance components: Mean square error equals estimated residual variance; mean square students minus estimated residual variance, difference divided by number of pseudo-forms equals estimated variance due to students; mean square pseudo-forms minus estimated residual variance, difference divided by number of students equals estimated variance due to pseudo-forms. For this pseudo-form combination, 97 percent of variance across the two sets of pseudo-form scores is explained by students. An estimate of how much score comparability is lost as a result of matrixing 20 percent of the test for these particular data is therefore just under three percent of the variation in scores. Of this variability, almost none is due to the pseudo-forms per se, an indication, in some regards, of the success of the scaling method in producing scores with equal means on the two sets of pseudo-form scores. This variability is, instead, reflected in the residual, also known as the error term.

Table 8 merits further explanation for two reasons. First, the estimated variance component for pseudo forms is nearly zero, rounding to .0001. Although this finding reflects the success of the scaling program in yielding score means that are equivalent across pseudo-forms, a contributing factor is the overlap in items across the two pseudo-forms. A near-zero estimated variance component is not at all problematic, except when it is negative, as was the case for the majority of pseudo-form pair comparisons. Because a negative variance component is not theoretically possible, the situation was handled following an established procedure of setting negative estimated variance components to zero after completing estimated variance component calculations (Brennan, 1983; as noted by Shavelson & Webb, 1991).

A second necessary explanation of Table 8 involves the Pearson correlation coefficient. As a more familiar statistical indication of the extent to which students scoring relatively high on one pseudo-form scored relatively high on the other, a correlation of .97 was observed between the two sets of scores in Table 8. Indeed, for nearly every data set studied (see summary statistics in Table 5), the correlation coefficient differed only nominally from the percent of variance accounted for by students, typically agreeing to the third place to the right of the decimal when both were expressed in decimal form.[3] This pattern generally held for all comparisons, including those where no items were in common across pseudo-forms. Therefore, the finding should not be considered an artifact of the overlap in items across pseudo-forms.

[3] The virtual equivalence of generalizability analysis and correlational results for these data is considered in detail in the Technical Note presented immediately following the reference section of this report.

IMPLICATIONS

Matrix sampling represents a likely tradeoff of comparability in student scores for breadth of items used to estimate school scores. The method employed above yields estimates of the error (specifically, noncomparability across students) introduced into student score comparisons by matrixing. Confronted with these data, a test developer or user might reasonably ask, how much error is it acceptable to add to student scores, relative to the case where all students take the same items? If, for example, the test carries high stakes related to high school graduation for individual students, it is likely that no additional error is acceptable. In cases without such high stakes for students, test developers and users should weigh likely consequences of a single-form test narrowing the curriculum against the possibility that matrixing may lead to the misinterpretation or misuse of an individual student's score. Indeed, with adequate data, test developers might also attempt to estimate the improvement in school score estimates obtained by matrix sampling at the student level for a particular test.

For testing programs where matrix sampling is viable, it is important to consider the number of items necessary to cover content adequately, in addition to the tolerance for noncomparability in student scores. Testing a broadly-defined content area will require a greater number of items than testing a narrowly defined one, as established by reference to curriculum frameworks and other statements of what students are supposed to know and be able to do by the time they are tested. If assessment results are in any way used for accountability purposes, concern for the likely impact on curriculum will suggest the need for a large number of test items as well. Regardless of the stakes attached to test scores, the consequences of potential misinterpretation of school-level results obtained on a too limited or inadequately representative sample of items should lead test developers and test users to prefer a large set of items to a small set, all other things being equal.

Once a test developer or user has established a likely number of items to be used in testing, the question of whether and to what degree matrix sampling should be employed must be framed in terms of available resources. Printing additional forms of the test, a necessity for matrixing in all but computerized testing environments, is costly and adds complexity to test administration, scoring, and data processing, as well as to interpretation of test results. With limited funds, a test developer or test user would choose to have fewer forms with a greater degree of matrixing, if the corresponding trade-off in student score comparability were acceptable. Any prevailing regulation requiring full disclosure of items after testing would add development costs on an annual basis. Regardless of whether matrixing is employed, significant resources will be needed to follow the preferred practice of replenishing the items on a test regularly and equating scores across test forms administered in successive years (Linn, 2000).

As for the specific procedures test developers and assessment staff might follow in designing matrix sampled tests, one approach is to pilot longer than necessary tests to allow analyses following the model described above to yield results that can be directly applied as estimates of the variability introduced by matrixing. Piloting a test twice as long as the final version, for example, would yield readily applicable estimates of variability for tests half as long and fully matrixed, as well as tests nearly as long as the pilot and only slightly matrixed. Of course, the possibility of student fatigue in

completing long tests would limit the applicability of the results to shortened test forms. In this case, researchers would need to monitor students for signs of fatigue during administration and examine responses for patterns suggestive of fatigue, such as the rate at which items are attempted, and for constructed-response questions, the consistency of answers and apparent effort over the course of the test. Piloting tests closer in length to the final version would lessen the impact of fatigue but also restrict analyses to a more limited range of items in common across pseudo-forms. Ideally, test developers might employ results from a study of items being used in test construction as a way of gauging likely test length and degree of matrixing that could be supported, piloting tests only as much longer than the final version as would be necessary to compute results for the proposed degree of matrixing.

When IRT scaling techniques are used to derive scores with negligible mean differences across matrixed or partially-matrixed pseudo-forms, it may be sufficient to treat correlations across such forms as alternate forms reliability estimates rather than incorporate generalizability analysis per se. As noted by Shavelson and Webb (1991, p. 94), "the reliability coefficient in classical theory is comparable to the generalizability coefficient for relative decisions" for generalizability studies with one (random) facet. In view of the relationship between generalizability coefficients and correlations of scores across pseudo-forms obtained with item response theory-based scaling, it appears both adequate and, certainly, more efficient computationally to rely upon correlations across pseudo-form scores, preferably in conjunction with measures such as RMSD.

LIMITATIONS AND EXTENSIONS

Our findings should be considered as particular to the test we studied. Although our work included both constructed-response and multiple-choice items in a variety of weighted combinations, we can only hope to have provided a rough indication of the range of results that might be obtained on other tests. Considering the importance of student score comparability and the possibilities for misinterpretation of an individual student's results when it is not known how much error they contain, assessment developers should strive to analyze their own field test item data to establish the degree of matrixing that corresponds approximately to a tolerable amount of error (specifically, noncomparability across students) in student scores. However, our study did not endeavor to shed light on what constitutes a tolerable amount of error in student scores, merely to quantify the amount of error that would be added through matrix sampling.

Our results reflect substantial methodological limitations. Among the more modest considerations is that raters may have affected findings for constructed-response scores. Rater reliability, which was not modeled in our generalizability analyses or elsewhere in this study, contributed to measurement error in scores. One imaginable way in which this might have happened is through a "halo" effect, whereby reading a well-reasoned answer to a constructed-response question predisposed the rater to attribute an undeservedly high mark to the next question answered by the same student. In our study, raters were a hidden or unidentified facet of generalizability analyses. See Brennan (2000) for a discussion of unidentified facets in generalizability studies.

Rater effects may have played into the more important limitation, that the method we used included measurement error in estimates of how comparable scores were across pseudo-forms. Students who missed an item during testing were counted as having missed it every time a pseudo-form score was generated. Somewhat less favorable correlations and RMSD results would have been expected if actual matrix sampled test forms had been created and administered to the same students on separate occasions.

Note that no equating of pseudo-forms took place in this study, only calibration. Because an operational testing program using matrix sampling would employ the statistical methodology of equating, it is possible that results obtained in our study might be improved upon, reducing the impact of matrixing.

In exploring the relationship of matrix sampling and score comparability, it may be useful to vary item positions in a manner that could allow for estimation of the contribution of context effects (Feldt & Forsyth, 1974; Kingston & Dorans, 1984) to loss in score comparability. Knowing which items appear most impervious to context effects would afford a more informed selection of items for use in matrix sampled tests. Test developers working with newly written items with unknown statistical properties should also examine model fit in using IRT scaling techniques, to ensure that no misperforming items are inadvertently included in estimates of score comparability lost to matrixing. This was not a particular concern in our study, as we used operational rather than unproved test items.

One way to extend these analyses would be to examine item content that contributes to score comparability across pairs of pseudo-forms. Although this could conceivably have been done with the current data, items included in this study were not designed to support systematic manipulation of multiple-choice or constructed-response characteristics likely to contribute to score comparability, such as the degree to which extended constructed-response items require explanation of one's answer. The items were also not written to afford a study of the comparability of scores on two forms built to cover particular content with several multiple-choice items versus a single constructed-response item. Thus, no effort was made to select pairs of pseudo-forms in this study to cover highly similar content or to vary the content overlap in a systematic manner. This limitation is particularly important, in that an especially well-designed study could contrast the statistical comparability of pairs of pseudo-form scores based on non-overlapping sets of items reflecting identical content coverage for each pseudo-form with results obtained for pairs with non-overlapping item sets reflecting systematically differing content coverage. The difference between such sets of results could provide guidance for test construction procedures designed to limit the degree of comparability lost by matrixing.

REFERENCES

Bock, R. D., & Mislevy, R. J. (1981). An item response curve model for matrix-sampling data: The California grade-three assessment. *New Directions in Testing and Measurement, 10,* 65-89.

Bock, R. D., & Mislevy, R. J. (1987). Comprehensive educational assessment for the states: The duplex design. *Evaluation Comment* (November 1987, pp. 1-16). Los Angeles: University of California at Los Angeles, Center for Research on Evaluation, Standards and Student Testing.

Bock, R. D., & Muraki, E. (1996). *Parscale.* Chicago: Scientific Software International.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City: The American College Testing Program.

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5-10.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues & Practices, 14,* 9-12, 27.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1995). Generalizability analysis for educational assessments. *Evaluation Comment* (Summer 1995, whole issue). Los Angeles: University of California at Los Angeles, Center for Research on Evaluation, Standards and Student Testing.

Feldt, L. S., & Forsyth, R. A. (1974). An examination of the context effect in item sampling. *Journal of Educational Measurement, 11,* 73-82.

Fitzpatrick, A. R., & Yen, W. M. (1999). Issues in linking *scores on alternative assessments: Effects of test length and sample size on test reliability and test equating.* Washington, DC: Council of Chief State School Officers.

Garg, R., Boss, M. W., & Carlson, J. E. (1986). A comparison of examinee sampling and multiple matrix sampling in test development. *Journal of Educational Measurement, 23,* 119-130.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In *Technical issues in large-scale performance assessment* (pp. 59-78; Report No. NCES 96-802). Washington, DC: U.S. Department of Education.

Kerins, T., & Brennan, M. (1987). Comments on "Comprehensive educational assessment for the states: The duplex design." *Evaluation Comment* (November 1987, pp. 18-19). Los Angeles: Center for Research on Evaluation, Standards and Student Testing, University of California at Los Angeles.

Kingston, N. M., and Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8,* 147-154.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Lord, F. M. (1962). Estimating norms by item sampling. *Educational and Psychological Measurement, 22,* 259-267.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133-161.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage Publications.

Yen, W. M. (1997). The technical quality of performance assessments: Standard errors of percents of students reaching standards. *Educational Measurement: Issues and Practice, 16,* 5-15.

TECHNICAL NOTE

The generalizability coefficient is equal to the variance component for students divided by the sum of the variance components for students and the residual. For Table 4, this is .6400/(.6400+.0178)=.9729. Because the variance component due to pseudo-forms is so very small, this yields the same result as the computation of the proportion of total variance attributable to students (.6400/(.6400+.0001+.0178)=.9729).

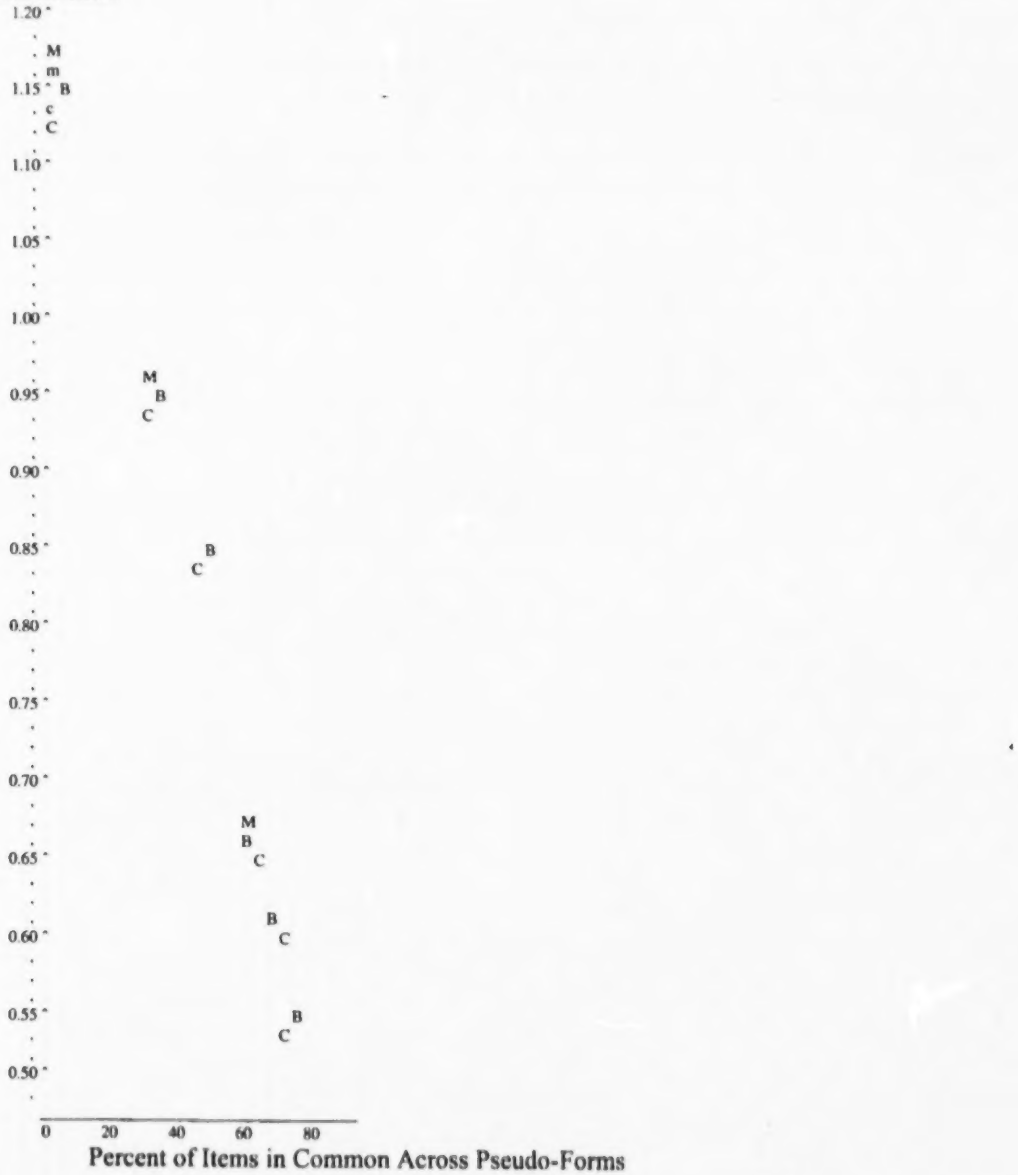
The generalizability coefficient reduces, through algebraic computations (and assuming that the number of pseudo-forms is 2) to [SS(p)-SS(pi)]/[SS(p)+SS(pi)] or, equivalently, to [SS(Total)-SS(i)-2SS(pi)]/[SS(Total)-SS(i)]. In computing the correlation between pseudo-form scores, we are not concerned about the SS(i); that is, we do not care if the average score on pseudo-form 1 is different from the average score on pseudo-form 2. All we care about are the differences between a person's score on pseudo-form 1 and the average for pseudo-form 1 and between the person's score on pseudo-form 2 and the average for pseudo-form 2. The denominator in the generalizability coefficient equation gives us the SS related to these score differences. The numerator is further reduced by twice the SS(pi). The SS(pi) can be thought of as the squared sum of the difference between each person's score on one pseudo-form and that person's average score across pseudo-forms and the difference between the average score for the pseudo-form and the overall average. Alternatively, it can be thought of as the squared sum of the difference between the person's score on one pseudo-form and the average score for the pseudo-form and the difference between the person's average score across pseudo-forms and the overall average.

The numerator of the generalizability coefficient is SS(p), which is the sum of squared differences between each person's average score and the overall average, minus SS(pi), the sum of squared sums of the difference between the person's score on one pseudo-form and the average score for the pseudo-form and the difference between the person's average score across pseudo-forms and the overall average. This reduces to twice the sum of squared differences between the person's average score and the overall average minus the sum of squared differences between the person's score and the average score for the pseudo-form. For these data, we know that the average score for each pseudo-form is very close to the overall average. Therefore, this reduces further to the sum of squared differences between the person's average score and the average score for the pseudo-form minus the sum of squared differences between the person's score on the pseudo-form and the person's average score.

Conceptually, for the generalizability coefficient to equal the correlation between pseudo-forms, the numerator of the generalizability coefficient must equal the sum of products of the difference between each person's score on pseudo-form 1 and the average score for pseudo-form 1, and the difference between each person's score on pseudo-form 2 and the average score for pseudo-form 2. When the average scores on the two pseudo-forms are virtually identical, this is very close to what we have.

FIGURE 1

Average Difference in Scores Between Pseudo Forms in Standard Error of Measurement Units (RMSD/SEM), Plotted Against Percent of Items in Common Across Pseudo-Forms



BEST COPY AVAILABLE

NOTE: 13 observations are hidden.

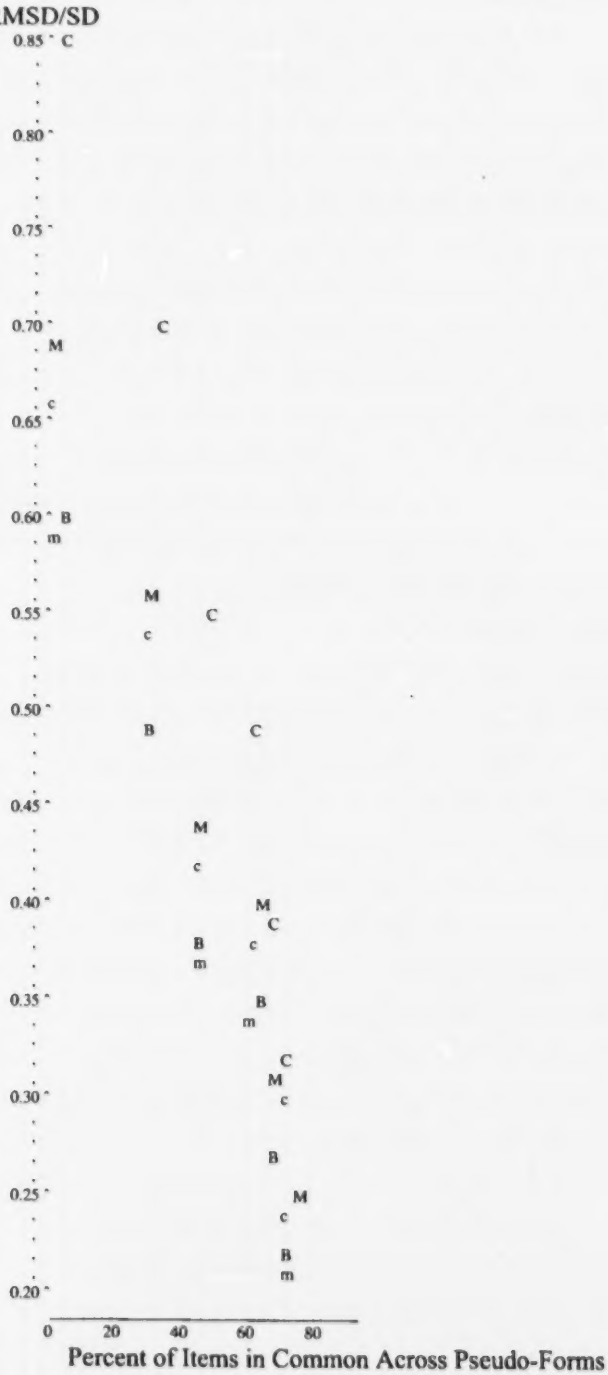
"C" indicates Constructed-response item results, based on pseudo-forms of three, four, or five 4-point items (many pseudo-forms of each).

"M" indicates Multiple-choice item results, based on several pseudo-forms of fifteen, twenty, or twenty-five items (many pseudo-forms of each).

"B" indicates results based on Both multiple-choice and constructed-response items, equally weighted; "c" and "m" reflect 2-to-1 weighting schemes favoring constructed-response and multiple-choice items, respectively.

FIGURE 2

Average Difference in Scores Between Pseudo-Forms in Standard Deviation Units (RMSD/SD), Plotted Against Percent of Items in Common Across Pseudo-Forms



26

BEST COPY AVAILABLE

NOTE: 2 observations are hidden.

"C" indicates Constructed-response item results, based on pseudo-forms of three, four, or five 4-point items (many pseudo-forms of each).

"M" indicates Multiple-choice item results, based on several pseudo-forms of fifteen, twenty, or twenty-five items (many pseudo-forms of each).

"B" indicates results based on Both multiple-choice and constructed-response items, equally weighted; "c" and "m" reflect 2-to-1 weighting schemes favoring constructed-response and multiple-choice items, respectively.

TABLE 1

Ways of Combining Six 5-Item Sets into Pseudo-Forms of 3, 4, or 5 Sets Each

	Item Sets Included in Each Form					
Form	**1**	**2**	**3**	**4**	**5**	**6**
Original*	**X**	**X**	**X**	**X**	**X**	**X**
Pseudo 3A	X	X	X			
Pseudo 3B	X	X		X		
Pseudo 3C	X	X			X	
Pseudo 3D	X	X				X
Pseudo 3E	X		X	X		
Pseudo 3F	X		X		X	
Pseudo 3G	X		X			X
Pseudo 3H	X			X	X	
Pseudo 3I	X			X		X
Pseudo 3J	X				X	X
Pseudo 3K		X	X	X		
Pseudo 3L		X	X		X	
Pseudo 3M		X	X			X
Pseudo 3N		X		X	X	
Pseudo 3O		X		X		X
Pseudo 3P		X			X	X
Pseudo 3Q			X	X	X	
Pseudo 3R			X	X		X
Pseudo 3S			X		X	X
Pseudo 3T				X	X	X
Pseudo 4A	X	X	X	X		
Pseudo 4B	X	X	X		X	
Pseudo 4C	X	X	X			X
Pseudo 4D	X	X		X	X	
Pseudo 4E	X	X		X		X
Pseudo 4F	X	X			X	X
Pseudo 4G	X		X	X	X	
Pseudo 4H	X		X	X		X
Pseudo 4I	X		X		X	X
Pseudo 4J	X			X	X	X
Pseudo 4K		X	X	X	X	
Pseudo 4L		X	X	X		X
Pseudo 4M		X	X		X	X
Pseudo 4N		X		X	X	X
Pseudo 4O			X	X	X	X
Pseudo 5A	X	X	X	X	X	
Pseudo 5B	X	X	X	X		X
Pseudo 5C	X	X	X		X	X
Pseudo 5D	X	X		X	X	X
Pseudo 5E	X		X	X	X	X
Pseudo 5F		X	X	X	X	X

* The student responded to all items on the original form. Each item set consists of one constructed-response and four multiple-choice items.

TABLE 2:

Median Summary Statistics for Original and Pseudo-Form Raw Scores, Computed Over 12 Test Forms of Multiple-Choice and Constructed-Response Items

Original or Pseudo-Form	Combined CR and MC			CR only			MC only			MC/CR Correlation
	Mean	SD	Coeff Alpha	Mean	SD	Coeff Alpha	Mean	SD	Coeff Alpha	
Original	**27.0**	**9.5**	**.88**	**11.6**	**4.9**	**.77**	**15.3**	**5.4**	**.86**	**.75**
3A	14.4	4.8	.79	6.7	2.5	.64	7.7	2.9	.75	.61
3B	14.3	5.0	.77	6.5	2.8	.59	7.8	2.9	.75	.62
3C	14.1	4.9	.80	6.4	2.4	.63	7.7	2.9	.76	.65
3D	13.8	4.8	.79	6.1	2.5	.62	7.5	2.9	.75	.64
3E	14.1	5.1	.76	6.2	2.9	.59	7.9	2.8	.74	.61
3F	13.8	4.9	.79	6.1	2.6	.63	7.9	2.8	.75	.63
3G	13.5	4.9	.78	5.8	2.6	.62	7.6	2.8	.74	.62
3H	13.8	5.2	.77	5.9	2.9	.59	8.0	2.8	.76	.64
3I	13.5	5.1	.76	5.6	2.9	.58	7.8	2.8	.75	.63
3J	13.2	5.0	.79	5.5	2.7	.64	7.7	2.9	.75	.65
3K	13.7	5.0	.77	6.1	2.8	.61	7.6	2.8	.74	.61
3L	13.4	4.8	.80	5.9	2.4	.65	7.6	2.9	.76	.65
3M	13.1	4.8	.79	5.7	2.5	.64	7.4	2.9	.75	.64
3N	13.5	5.1	.78	5.7	2.8	.62	7.7	2.9	.76	.64
3O	13.0	5.0	.78	5.5	2.8	.62	7.4	2.9	.75	.63
3P	12.9	4.9	.81	5.3	2.6	.68	7.5	2.9	.76	.68
3Q	13.2	5.2	.77	5.4	2.9	.61	7.8	2.8	.75	.63
3R	12.8	5.1	.77	5.2	2.9	.61	7.5	2.8	.74	.63
3S	12.7	5.0	.80	5.0	2.6	.67	7.7	2.9	.76	.65
3T	12.6	5.3	.78	4.9	3.0	.64	7.6	2.9	.76	.65
4A	18.8	6.5	.82	8.5	3.5	.67	10.3	3.7	.79	.66
4B	18.6	6.3	.84	8.3	3.1	.70	10.3	3.7	.80	.69
4C	18.2	6.2	.83	8.1	3.2	.69	10.0	3.7	.80	.67
4D	18.6	6.6	.83	8.2	3.4	.67	10.4	3.7	.80	.69
4E	18.2	6.5	.82	7.9	3.5	.67	10.2	3.7	.80	.68
4F	18.0	6.3	.84	7.7	3.3	.70	10.2	3.7	.80	.70
4G	18.4	6.6	.82	7.9	3.6	.67	10.5	3.7	.80	.68
4H	18.0	6.5	.82	7.6	3.5	.67	10.3	3.6	.80	.67
4I	17.8	6.5	.83	7.4	3.4	.70	10.2	3.7	.80	.69
4J	17.7	6.7	.82	7.3	3.7	.68	10.4	3.7	.80	.69
4K	17.8	6.6	.83	7.7	3.4	.69	10.3	3.7	.80	.68
4L	17.5	6.5	.82	7.5	3.4	.68	9.9	3.7	.80	.68
4M	17.4	6.3	.84	7.3	3.3	.72	10.0	3.8	.80	.70
4N	17.2	6.6	.83	7.2	3.6	.70	10.0	3.7	.80	.70
4O	17.1	6.7	.83	6.8	3.7	.70	10.2	3.7	.80	.69
5A	23.1	8.0	.86	10.1	4.1	.73	12.9	4.5	.83	.71
5B	22.6	7.9	.86	9.9	4.1	.72	12.6	4.5	.83	.71
5C	22.5	7.8	.87	9.7	4.0	.75	12.7	4.6	.83	.73
5D	22.5	8.1	.86	9.6	4.3	.73	12.8	4.5	.84	.73
5E	22.2	8.2	.85	9.3	4.3	.73	12.8	4.5	.83	.72
5F	21.8	8.1	.86	9.1	4.2	.74	12.6	4.6	.83	.72

BEST COPY AVAILABLE

TABLE 3:

Median Standard Deviation (SD) and Standard Error (SE) for Item Response Theory Based Scale Scores on Original and Pseudo-Forms

Original or Pseudo-Form	Weight of Constructed-Response(CR) and Multiple-Choice(MC) Scores										Corr. of MC w/ CR
	CR Only		MC Only		.50CR/.50MC		.33CR/.67MC		.67CR/.33MC		
	SD	SE	SD	SE	SD	SE	SD	SE	SD	SE	
Original	**.88**	**.47**	**.92**	**.39**	**.85**	**.30**	**.86**	**.30**	**.85**	**.34**	**.76**
3A	.80	.60	.87	.51	.75	.39	.77	.39	.75	.44	.62
3B	.78	.62	.87	.51	.75	.40	.77	.40	.74	.45	.62
3C	.80	.59	.87	.50	.76	.39	.78	.39	.76	.43	.66
3D	.80	.59	.87	.50	.76	.39	.78	.39	.75	.44	.65
3E	.78	.63	.86	.51	.74	.41	.76	.40	.73	.46	.61
3F	.80	.61	.87	.51	.75	.40	.77	.39	.75	.44	.65
3G	.79	.61	.87	.51	.75	.40	.77	.39	.75	.44	.63
3H	.79	.61	.86	.51	.75	.40	.77	.40	.74	.44	.65
3I	.78	.62	.87	.51	.75	.40	.77	.40	.74	.45	.65
3J	.81	.59	.87	.50	.76	.39	.78	.39	.76	.43	.67
3K	.80	.60	.87	.50	.75	.39	.77	.39	.75	.44	.62
3L	.81	.58	.87	.50	.76	.38	.79	.39	.76	.43	.66
3M	.81	.58	.87	.50	.76	.39	.78	.39	.76	.43	.66
3N	.81	.59	.87	.50	.76	.39	.79	.39	.76	.43	.66
3O	.80	.59	.87	.50	.76	.39	.79	.39	.76	.43	.66
3P	.82	.57	.87	.50	.78	.38	.80	.38	.78	.41	.69
3Q	.79	.61	.86	.51	.75	.40	.78	.39	.75	.44	.65
3R	.78	.61	.87	.51	.75	.40	.77	.40	.75	.44	.65
3S	.81	.58	.87	.50	.77	.39	.79	.39	.76	.43	.68
3T	.81	.59	.87	.51	.77	.39	.79	.39	.77	.43	.68
4A	.83	.56	.89	.46	.79	.36	.81	.36	.78	.40	.66
4B	.84	.54	.89	.46	.80	.35	.82	.35	.80	.39	.70
4C	.84	.54	.89	.45	.79	.35	.82	.35	.79	.39	.69
4D	.83	.54	.89	.45	.80	.36	.82	.35	.79	.40	.70
4E	.83	.55	.89	.45	.79	.36	.81	.35	.79	.40	.69
4F	.85	.53	.89	.45	.81	.35	.82	.35	.81	.38	.72
4G	.83	.56	.89	.46	.79	.36	.81	.36	.79	.40	.69
4H	.82	.56	.89	.46	.79	.36	.81	.36	.78	.40	.69
4I	.84	.54	.90	.45	.80	.36	.82	.35	.80	.39	.71
4J	.83	.55	.89	.45	.80	.36	.81	.35	.79	.40	.71
4K	.84	.54	.89	.45	.80	.35	.82	.35	.79	.39	.70
4L	.84	.54	.89	.46	.80	.36	.81	.35	.79	.39	.70
4M	.85	.52	.90	.45	.81	.35	.83	.35	.81	.38	.72
4N	.85	.53	.89	.45	.81	.35	.82	.35	.81	.39	.72
4O	.84	.54	.89	.46	.80	.36	.82	.35	.80	.40	.71
5A	.86	.50	.91	.42	.82	.33	.84	.33	.82	.36	.73
5B	.86	.51	.91	.42	.82	.33	.84	.33	.82	.37	.72
5C	.87	.49	.91	.41	.83	.32	.85	.32	.83	.35	.74
5D	.87	.50	.91	.42	.83	.32	.84	.32	.83	.36	.75
5E	.86	.51	.91	.42	.82	.33	.84	.33	.82	.37	.74
5F	.87	.49	.91	.41	.83	.32	.85	.32	.83	.36	.75

Note: Item Response Theory-based scaling yields mean scale scores of approximately zero on the original form and each pseudo-form. Because the means all round to .00, they are not presented in the table above.

TABLE 4:

Pseudo-Form Lengths Employed to Study Various Combinations of Multiple-Choice and Constructed-Response Items, Together and Separately

Study Factors		Pseudo-Form Length, by Item Types Included					
		Combined CR and MC		Constructed-Response		Multiple-Choice	
Percentage of Items and Item Sets in Common Across Pseudo-Forms	Number of Pairs of Pseudo-Forms Compared in Study*	Items	Maximum Total Score Points	Items	Maximum Total Score Points	Items	Maximum Total Score Points
80	15	25	40	5	20	20	20
75	32	20	32	4	16	16	16
67	18	15	24	3	12	12	12
50	45	20	32	4	16	16	16
33	18	15	24	3	12	12	12
0	10	15	24	3	12	12	12

*Note: Comparisons were replicated on each of twelve forms used in the testing program, such that twelve times as many replications as listed above were performed on each pair of pseudo-forms compared in this study.

TABLE 5:

Median Root Mean Squared Difference and Pearson Correlation Between Pseudo-Form Scores, with Reliability Estimates and Number of Replications

Score	Percent of Items in Common	Number of Replications	RMSD	Reliability of Pseudo-Form Scores	Correlation Between Pseudo-Form Scores
MC(.5)/CR(.5) Composite	0%	120	.45	.73	.82
MC(.5)/CR(.5) Composite	33%	204	.37	.74	.88
MC(.5)/CR(.5) Composite	50%	540	.30	.80	.93
MC(.5)/CR(.5) Composite	67%	204	.26	.73	.94
MC(.5)/CR(.5) Composite	75%	384	.22	.80	.96
MC(.5)/CR(.5) Composite	80%	180	.18	.84	.98
CR Only	0%	120	.67	.44	.64
CR Only	33%	204	.56	.45	.76
CR Only	50%	540	.45	.58	.85
CR Only	67%	204	.39	.44	.88
CR Only	75%	384	.32	.57	.92
CR Only	80%	180	.27	.67	.95
MC Only	0%	120	.59	.66	.77
MC Only	33%	204	.49	.67	.84
MC Only	50%	540	.39	.74	.91
MC Only	67%	204	.34	.67	.92
MC Only	75%	384	.28	.74	.95
MC Only	80%	180	.23	.79	.97
MC(.33)/CR(.67) Composite	0%	120	.49	.67	.78
MC(.33)/CR(.67) Composite	33%	204	.41	.67	.85
MC(.33)/CR(.67) Composite	50%	540	.33	.76	.91
MC(.33)/CR(.67) Composite	67%	204	.29	.67	.93
MC(.33)/CR(.67) Composite	75%	384	.24	.75	.96
MC(.33)/CR(.67) Composite	80%	180	.20	.81	.97
MC(.67)/CR(.33) Composite	0%	120	.46	.75	.83
MC(.67)/CR(.33) Composite	33%	204	.38	.75	.88
MC(.67)/CR(.33) Composite	50%	540	.30	.81	.93
MC(.67)/CR(.33) Composite	67%	204	.27	.75	.94
MC(.67)/CR(.33) Composite	75%	384	.22	.81	.96
MC(.67)/CR(.33) Composite	80%	180	.18	.85	.98

BEST COPY AVAILABLE

TABLE 6

Average Difference in Scores Between Pseudo-Forms in Standard Error of Measurement Units (RMSD/SEM), by Percent of Items in Common Across Pseudo-Forms

	Percent of Items in Common Across Pseudo-Forms					
Score	**0%**	**33%**	**50%**	**67%**	**75%**	**80%**
CR Only	1.13	.93	.84	.65	.60	.54
MC(.33)/CR(.67) Composite	1.14	.94	.84	.66	.60	.55
MC(.5)/CR(.5) Composite	1.15	.94	.85	.67	.61	.55
MC(.67)/CR(.33) Composite	1.17	.96	.85	.67	.61	.55
MC Only	1.17	.96	.85	.67	.61	.55

TABLE 7:

Average Difference in Scores Between Pseudo-Forms in Standard Deviation Units (RMSD/SD), by Percent of Items in Common Across Pseudo-Forms

	Percent of Items in Common Across Pseudo-Forms					
Score	**0%**	**33%**	**50%**	**67%**	**75%**	**80%**
CR Only	.85	.70	.55	.49	.39	.32
MC Only	.69	.56	.44	.40	.31	.25
MC(.33)/CR(.67) Composite	.66	.54	.42	.38	.30	.24
MC(.5)/CR(.5) Composite	.60	.49	.38	.35	.27	.22
MC(.67)/CR(.33) Composite	.59	.48	.37	.34	.27	.21

TABLE 8:

Example of Generalizability Study Results, Form 1, Design With 80 Percent of Items in Common Across Pseudo-Forms (500 Students, 2 Scores per Student)

Source of Variation	df	Sums of Squares	Mean Squares	Estimated Variance Components	Percentage of Total Variance Accounted For
Students	499	647.6108	1.2978	0.6400	97.29
Pseudo-Forms	1	0.0675	0.0675	0.0001	0.02
Residual	499	8.8588	0.0178	0.0178	2.70

APPENDIX A

Percent of Items Covering Content Across the Twelve Original Test Forms, By Constructed-Response versus Multiple-Choice Item Format

	MC	CR	Total
Measurement	13%	25%	16%
Change	7%	14%	9%
Structure	12%	0%	9%
Data	23%	25%	24%
Number	12%	18%	13%
Procedures	21%	7%	19%
Space/Dimension	12%	11%	11%
Number of items	**112**	**28**	**140**

Note: Only the content category noted as primary for each item is indicated above. Most constructed-response items addressed a secondary content area, as did about one-fifth of multiple-choice items.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").